یونی کد چیست؟

مقدمه:

از چند سال پیش در کشورمان، استفاده از کامپیوتر با سرعت سرسام‌آوری جای خود را در تمامی عرصه‌ها باز کرد و سیل کامپیوترهای شخصی و تجهیزات جانبی آنها به سوی کشور سرازیر شد. اما بایستی اعتراف کرد که با وجود این که سرعت سوق به سوی تکنولوژی دیجیتال در ایران روند خوبی را طی نموده، اما در زمینه ارایه اطلاعات و پردازش آن به زبان فارسی تشتتی در این میان ایجاد گردید. یکی از عوامل موثر در این ناهماهنگی، نبود الگویی واحد برای ذخیره و پردازش و نمایش اطلاعات بر روی رسانه‌های جدید اطلاع‌رسانی همچون کامپیوتر در سطت ملی است .

نرم‌افزارهای متفاوت، با فرمت‌های مختلف، کدهای فارسی گوناگون و ... در حال استفاده‌اند و روزانه میزان قابل توجهی از اطلاعات را در خود جای می‌دهند. اگر از آن دسته از مراکزی که به دلیل عدم آگاهی کافی اطلاعات را به صورت ناقص جمع‌آوری و وارد می‌کنند (که حدود ۸۰ درصد جامعه مورد نظر را تشکیل می‌دهند) بگذریم به تفرق، اختلاف و اعمال سلیقه‌های مختلف در سایر مراکز خواهیم رسید که برای نمونه به اختلاف در مورد کدنویسه‌های به کار رفته برای حروف فارسی روی کامپیوتر می‌توان اشاره کرد .

نتیجه ادامه روند جاری

در مورد مراکزی که به هر حال مشغول سرمایه‌گذاری در بخش ورود، پردازش و نمایش اطلاعات هستند مسیله به نوع دیگری خود را نشان خواهد داد. این گونه مراکز تا زمانی که پای خود را از محدوده مرکز خود فراتر نگذاشته‌اند مشکلی نخواهند داشت، ولی به محض آنکه بخواهند با مراکز اطلاعاتی و تحقیقاتی دیگر ارتباط برقرار کرده یا به مبادله اطلاعات با این مرکز بپردازند متوجه خواهند شد که سال‌ها سرمایه‌های خود را بر باد داده‌اند.

همین مشکل در سطح ملی برای ایجاد یک مرکز اطلاعات ملی رخ خواهد نمود. زمانی این مشکل ملی بیشتر نمود پیدا می‌کند که بحث شبکه جهانی اینترنت نیز به میان آید.

اینترنت به عنوان کلیدی برای ارتباط با دیگر مراکز اطلاعاتی - به علت در دسترس بودن آسان و همچنین حجم عظیم اطلاعات موجود در آن- یکی از مهم‌ترین موضوعاتی خواهد بود که به علت عدم وجود یک سیستم جهانی برای ذخیره، بازیابی، پردازش و نمایش اطلاعات و به طور کلی مبادله اطلاعات که جنبه‌های ملی نیز داشته باشد، دارای نقاط ضعفی است که ما را از بهره‌برداری مناسب در جهت منافعمان باز می‌دارد .

راه حل چیست؟

از زمانی که اولین گزارش «زبان فارسی و کامپیوتر» در سال ۱۳۵۶ در دانشکده ریاضی و کامپیوتر دانشگاه صنعتی شریف ارایه شد، تا امروز که شبکه اینترنت چهره دیگری به اطلاع‌رسانی داده است، مدت زیادی می‌گذرد. امروزه دیگر محدودیت‌های سخت‌افزاری یا نرم‌افزاری نمی‌تواند مانع پیاده‌سازی یک سیستم ذخیره‌سازی، نمایش، و تبادل اطلاعات چندزبانه گردد. امروزه مؤسسات بزرگ استانداردسازی چون ایزو (ISO) وW ۳ Consortium نیز، در استانداردهایشان مشکلات و مسایل مربوط به جهانی‌سازی را در نظر می‌گیرند تا امر تبادل اطلاعات چند زبانه را تسهیل نمایند. اما به نظر می‌رسد که به دلیل عدم تضور ایرانیان و فارسی‌زبان‌ها در این روند، زبان

فارسی قدری غریب مانده و کمتر به آن توجه شده است. به عنوان مثال، هنوز در بین صدها مجموعه‌نویسه (Character Set)ثبت شده در اینترنت توسط یانا (Internet Assigned Number Authority)، تنها یک مجموعه‌نویسه ثبت شده متعلق به زبان فارسی است که آن هم کد پیج اختصاصی شرکت آیبیام است. حتی در مورد استاندارد کلی تبادل اطلاعات نیز قالبی که مورد توافق همه باشد وجود ندارد. سه قالب موجود، ایران سیستم، استاندارد ۲۹۰۰ و استاندارد ۳۳۴۲ ، هر یک ایراداتی دارند که سبب شده است شرکت‌ها و مؤسسات داخلی به جدول‌های خاص خود روی آورند تا بتوانند نیازهای خود را تا حدی رفع سازند.

اخیراً راه‌حل‌هایی در هر یک از مسایل خاص مربوط به تبادل اطلاعات برای بین‌المللی‌سازی در نظر گرفته شده است که با وجود این که این موارد کامل‌تر از جداولی است که در ایران برای حل مشکلات تبادل اطلاعات زبان فارسی ایجاد گردیده، ولی به خاطر عدم وجود مراجع موثق در مورد خط و زبان فارسی برای استانداردگذاران، مسایل خاص این زبان یا در نظر گرفته نشده و یا به شکل ناقص منظور شده است. خوشبختانه بسیاری از این استانداردها امکان گسترش بعدی را در نظر گرفته‌اند که روند تصحیح را تسهیل می‌کند.

یونی‌کد چیست؟

از جمله استانداردهای بین‌المللی که کامل‌تر از بقیه استانداردهای موجود به رفع نیازهای مربوط به تبادل اطلاعات چندزبانه پرداخته‌است، می‌توان به استاندارد یونی‌کد اشاره کرد.

این استاندارد، تقریباً توسط تمامی شرکت‌های بین‌المللی کامپیوتری، مانند آیبیام، مایکروسافت، و سان، و نیز موسسات ملی استاندارد در کشورهای مختلف جهان برای تبادل اطلاعات چندزبانه مورد توافق قرار گرفته است و سرعت رشد بسیار زیادی نیز در میان کاربران دارد. همین‌طور، در حال حاضر کلیه استانداردهای جدیدی که برای شبکه اینترنت طراحی می‌شوند، این دو استاندارد را به‌عنوان کدپیج پیش‌فرض می‌پذیرند که استاندارد XML و زبان جاوا از آن جمله‌اند.

به زبان ساده می‌توان گفت که یونی‌کد روشی برای تبدیل متون به رشته‌های عددی قابل ذخیره در کامپیوتر است. روش‌های گوناگونی برای این کار وجود دارند، ولی مزیت یونی‌کد نسبت به آنها، این است که یک روش کامل جهانی است؛ به این معنی که حروف همه زبان‌های دنیا و تمامی علایم مورد استفاده همه مردم جهان در آن آمده‌اند و همچنین در همه‌جا قابل نمایش است و نیاز به امکانات خاصی ندارد. البته یونی‌کد هنوز جوان است ولی امروزه بسیاری نرم‌افزارهای رایج در جهان (از جمله همه مرورگرهای جدید اینترنت) آن را پشتیبانی می‌کنند.

از مهم‌ترین مزایایی که یونی‌کد برای زبان فارسی دارد (مثل بسیاری زبان‌های دیگر) می‌توان موارد زیر را نام برد:

۱ .در نسخه استاندارد هر نرم‌افزاری که از این استاندارد پشتیبانی کند، می‌توان فارسی نوشت یا متون فارسی را خواند. بدین ترتیب دیگر نیازی به تأمین نسخه‌های خاص فارسی یا عربی نیست.

۲ .برای خواندن متون فارسی که توسط شرکت خاصی نوشته شده‌اند، نیازی به داشتن فونت خاص آن شرکت نداریم و هر متن فارسی که با استاندارد یونی‌کد، کدگذاری شده باشد، با هر فونت یونی‌کدی قابل مشاهده است.

۳ .امکان استفاده همزمان از زبان‌های فارسی و انگلیسی را تأمین می‌کند.

۴ .بدون استفاده از فونت‌های خاص امکان استفاده از علایم خاص را فراهم می‌کند.

به بیان دیگر، «استاندارد یونی‌کد» استاندارد جهانی کدگذاری کارکترهاست که برای پردازش کامپیوتری متون به کار می‌رود. این استاندارد همان کاراکترها و کدهای استاندارد ISO/IEC ۱۰۶۴۶ را داراست و کاملا با آن سازگار است. پس در واقع هر پیاده‌سازی سازگار با یونی‌کد، با ISO/IEC۱۰۶۴۶ نیز سازگار است.

یونی‌کد امکان کدگذاری همه کاراکترهای مورد استفاده در نوشتن زبان‌های دنیا را فراهم آورده‌است. این استاندارد از کدگذاری ۱۶بیتی استفاده می‌کند که برای بیش از۶۵۰۰۰ نویسه (کاراکتر) جا فراهم می‌کند. اگر چه ۶۵۰۰۰ نویسه برای کدگذاری اکثر نویسه‌هایی که در زبان‌های مهم دنیا استفاده می‌شود کافی است، با این حال یونی‌کد شیوه‌گسترشی به‌نام UTF-۱۶ فراهم‌کرده‌است که امکان اضافه‌کردن حدود یک میلیون نویسه دیگر را نیز می‌دهد. این دامنه برای کلیه نویسه‌های عالم، از جمله پوشش کامل همه خط‌های باستانی (همچون خط میخی) نیز کافی است.

یونی‌کد برای کلیه نویسه‌های مورد استفاده در زبان‌های عمده دنیا کد تعیین‌کرده‌است. به‌علت گسترده‌بودن فضای تخصیص نویسه، این استاندارد بسیاری از نمادهای لازم برای حروف‌چینی را نیز در بر گرفته‌است. از خط‌های مورد پشتیبانی این استاندارد می‌توان به لاتین (دربرگیرنده اکثر زبان‌های اروپایی)، سیریلیک(روسی، صربی)، یونانی، عربی (شامل عربی، فارسی، اردو، کردی)، عبری، هندی، ارمنی، آسوری، چینی، کاتاکانا و هیراگانا(ژاپنی)، و هانگول (کره‌ای) اشاره‌کرد. به‌علاوه، تعداد زیادی نماد ریاضی و فنی علایم نقطه‌گذاری، پیکان، و علامت‌های متفرقه‌در این استاندارد وجود دارد. این استاندارد برای علامت‌های ترکیب‌شونده یا اعراب‌ها نیز کدهایی در نظر گرفته‌است که‌از جمله آنها علامت‌هایی چون «?» (مد)هستند که در ترکیب حروف پایه، حروف تغییرلحن یافته‌ای چون «?» را می‌سازند.

به طور کلی، بعضی از مشخصات یونی‌کد به شرت زیر است:

نویسه‌های شانزده‌بیتی
یکی‌سازی (اختصاص یک کد به نویسه‌های مشترک در چند زبان مختلف(
نویسه، نه شکل (یک «ع»، و نه چهارتا: «ع»، «ع»، «ع»، «ع(«
بار معنایی (حرف‌بودن، مقدار عددی، (...

در استاندارد یونی‌کد، نویسه‌های فارسی در بلوک مربوط به‌خط عربی قرار دارند .این بلوک‌برای دربرگرفتن نویسه‌های زبان‌هایی که‌از خط عربی‌استفاده‌می‌کنند، مثل فارسی، اردو، پشتو، هندی، و کردی گسترش یافته است. این بلوک نشانه‌های قرآنی از قبیل نشانه‌های سجده و پایان آیه، و علایم وقف را نیز در بردارد.

در یونی‌کد با وجود یکی‌سازی کدهای حروف مشترک، برای حروف فارسی که بار معنایی یا نمایشی متفاوت با حروف عربی دارند، نویسه‌های جداگانه درنظر گرفته‌شده‌است. یعنی کلیه حروف خاص فارسی (پ، چ، ژ، گ) و نیز «ک» و «ی» فارسی که‌با حرف مشابه‌در عربی‌تفاوت نمایشی دارند، مکان‌جداگانه‌ای به‌خود اختصاص داده‌اند. کلیه اعراب‌های متداول حضور دارند و میان شکل‌فارسی/اردو و عربی ارقام نیز به‌علت شکل و رفتار متفاوت، تفاوت‌هایی منظور گشته‌است.

از طرف دیگر، علایم نقطه‌گذاری چون نقطه و فاصله که‌شکل یکسانی‌در خط‌های لاتین و عربی دارند، کد یکسان دارند. علایمی‌چون پرانتز نیز، بسته به جهت متن، آینه‌ای می‌شوند، به طور مثال، نویسه ۰۰۲۸ نماینده «پرانتز باز» است، و نه«پرانتز سمت‌چپ .«یونی‌کد اتصال‌مجازی و فاصله مجازی را نیز تحت‌نام‌های «اتصال با عرض‌صفر» و» بی‌اتصالی با عرض‌صفر» به‌رسمیت می‌شناسد.

بدن ترتیب ملاحظه می‌شود که برای حل مشکلات موجود، و نیز رفتن به سوی یک استاندارد مقبول و همه‌جانبه، استاندارد یونی‌کد، روشی مناسب به نظر می‌رسد.

اصطلاحات:

نویسه: در مقابل .character کوچک‌ترین واحد متن. مثلاً یک حرف لاتین، یک اعراب فارسی، یکی علامت نقطه‌گذاری، یک نشانه بریل، یا یک نماد ریاضی

شکل: در مقابل .glyph کوچک‌ترین واحد نمایش متن. برای بعضی نویسه‌ها مثل حروف فارسی و هندی ممکن است چند شکل موجود باشد. مثلاً « ب » و « ع » از اشکال‌نمایشی محسوب می‌شوند

مجموعه‌نویسه: در مقابل .character setمجموعه‌ای از نویسه‌ها که‌به‌هر نویسه عددی اختصاص‌می‌دهد که نماینده آن نویسه محسوب می‌شود و در تبادل اطلاعات مورد استفاده‌قرار می‌گیرد

مجموعه کد: در مقابل .codepage سیستمی که به‌هر نویسه دنباله مشخصی‌از بایت‌ها را متناظر می‌کند. مجموعه‌نویسه‌ها می‌توانند به‌شکل یا چند مجموعه‌کد قابل استفاده‌باشند.

وبلاگ کرانه گمنام **http://www.denaboy.persianblog.ir**

وبلاگ آموزش های کاربردی **http://www.et4ir.blogfa.com**

وبسایت فرزند دنا **http://dena2.coo.ir**

==========

مدیر سایت و وبلاگ ها : *فرید نیک اقبالی*